Steven Layne sol307
Amanuel Alemu aea592
Zavier Henry zrh561

What feature(s) did you try?

- Confounding Box Area
- Confounding Box Aspect Ratio
- Confounding Box Width
- Confounding Box Height
- Curvature

Were they continuous or discrete?

We tried to implement them all as continuous or discrete. However, We struggled to get any luck with continuous classifications. Whenever we attempted to make variables continuous it always resulted in a classification that was much worse than the basic classification. So we chose to discretize everything as that either resulted in Classifications that mirrored the basic classification or improved it by a slight amount.

How did you determine thresholds for discrete features?

To determine thresholds we found the average, standard deviation, median , maximum, and minimum values that appeared for Box Area, Aspect ratio, width , and height. From there we tried to play around with boundaries by making neutral zones which spanned within the standard deviation and then made separate classifications for being at either side of the standard deviation. From there we moved numbers around slightly and did weighting a bit differently until we got a flavor of the variables that we felt comfortable with.

How well did it work?

The only classifier we were able to get a lot of success out of was the height classifier. We were able to increase the drawing accuracy from 60 percent to 80 percent while the text accuracy stayed hovering around 40 percent.

Unfortunately we could not get it all to work as well as we would have liked. We spent most of the time during the assignment trying to figure out how to discretize the variables in a way that improved the classifier however, our lack of understanding of distributions may have been a limiting factor as we could not find a method that worked that well.

From our understanding the general issue that surrounded around this was that since the continuous variables work on the gaussian distribution and some of the values cannot be negative. Using a gaussian distribution did not necessarily make the most sense as values cannot exist below zero in things like box area, aspect ratio, width, and height. In an essence the data does not come from a normal distribution.

Note: The following Confusion Matrices were generated using the macro average of a 10 cross fold validation. This validation can be run from calling .classify('trainingFilesDirectory', numFolds) within the stroke labeler class.

Basic HMM Classification Results
Confusion Matrix

True Label	Classified as Drawing	Classified as Text	Percent Correct
Drawing	833.0	441.0	65.38 %
Text	431.0	310.0	41.84 %

{'text': {'text': 310.0, 'drawing': 431.0}, 'drawing': {'text': 441.0, 'drawing': 833.0}}

HMM Classification Results
Confusion Matrix

True Label	Classified as Drawing	Classified as Text	Percent Correct
Drawing	910.0	364.0	71.43 %
Text	501.0	240.0	32.39 %

{'text': {'text': 240.0, 'drawing': 501.0}, 'drawing': {'text': 364.0, 'drawing': 910.0}}